Metadata for Language Resources
The ISO-Cat Component Model

Daan Broeder & Peter Wittenburg
Max Planck Institute for Psycholinguistics

Metadata for Language Resources

- MPI for Psycholinguistics since 2000 active:
  - Metadata set design: IMDI
  - Metadata tools development
  - Archiving software building on IMDI
  - International metadata related projects: ISLE, ECHO, INTERA, DAM-LR, CLARIN
- Currently our catalog contains 90000 metadata descriptions whereof 30000 harvested from other organizations

Lessons and Observations

- LR Metadata sets seem to have stabilized but long term support must be assured.
- There is no single best suited set or schema:
  - Depends on granularity, purpose, available time, history, affiliations
- Coverage although grown is still limited compared to number of resources
- Having (sub-)discipline compliant terminology is crucial
- Still limited willingness to create “rich” metadata
  - Much work (for the benefit of others)
  - Only now we see applications where depositors profit from MD
  - Need also pressure from funding agencies

EU CLARIN Project

- Make LR and LRT available and usable to language and SSH researchers
- Create federation type infrastructure
  - Authentication & Authorization
  - Persistent Identifiers for resources
  - Workflow management
  - Metadata infrastructure for resources and tools
    - creation, harvesting, exploitation

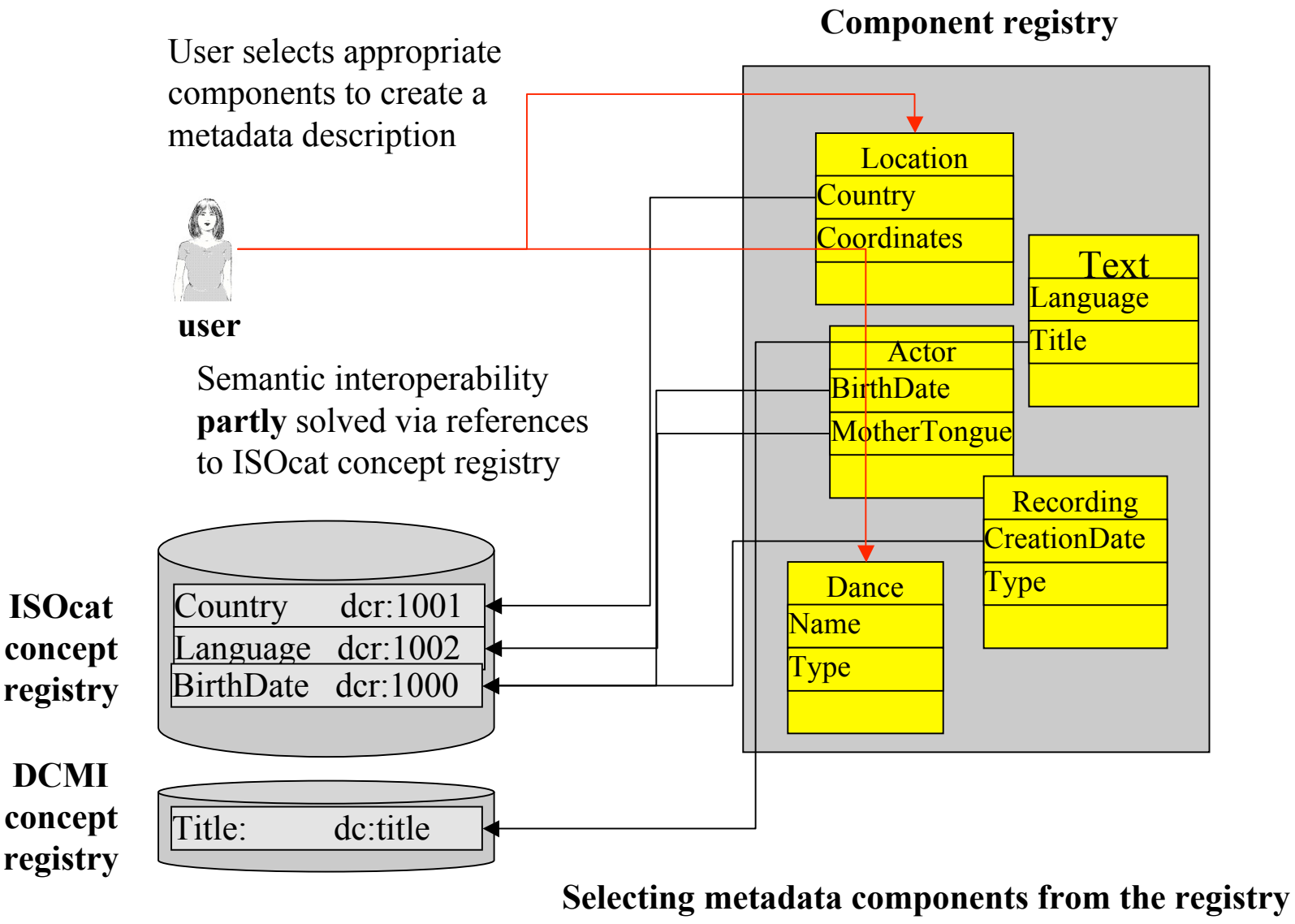
CLARIN Metadata Infrastructure

- Not a new metadata set.
- But an infrastructure aimed at reuse of existing metadata schema & elements
- Allowing “peaceful coexistence”.
  - of big installed bases: TEI, IMDI, OLAC, …
  - of resource creators' own schemas provided these
    - have explicit semantics
    - reuse existing elements where possible
    - and are registered

ISO Data Category Registry

- Registry of linguistic concepts, human & machine accessible
- Model and content maintained by ISO-TC37
  - On-line start of 2009
- Only Concepts, no relations
  - avoid theory bias
- Already many linguistic concepts inside
  - Content depends on community input
  - and will be subject to board review
- Secure persistence of metadata concept definitions

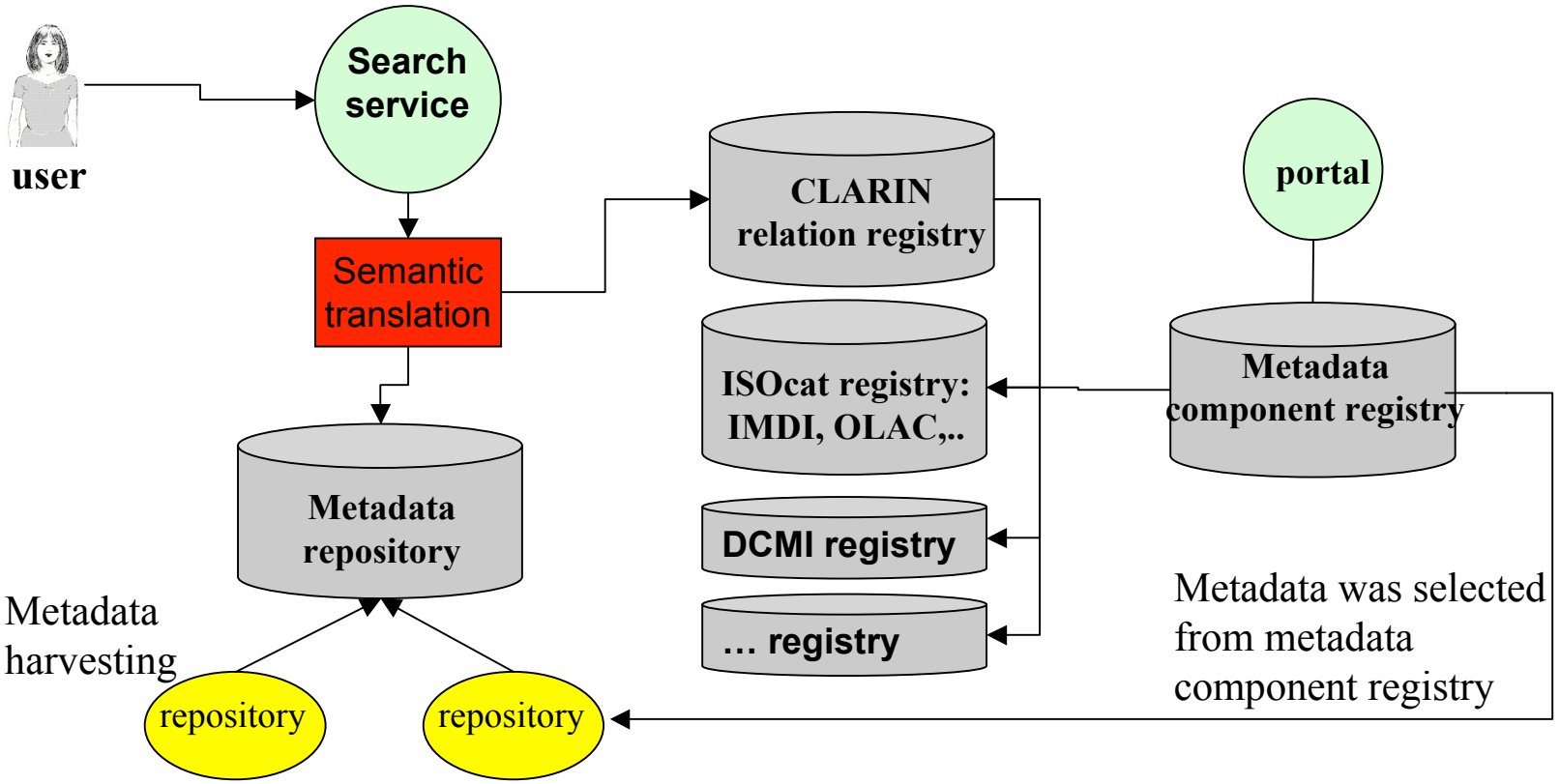
CLARIN Metadata



CLARIN Metadata

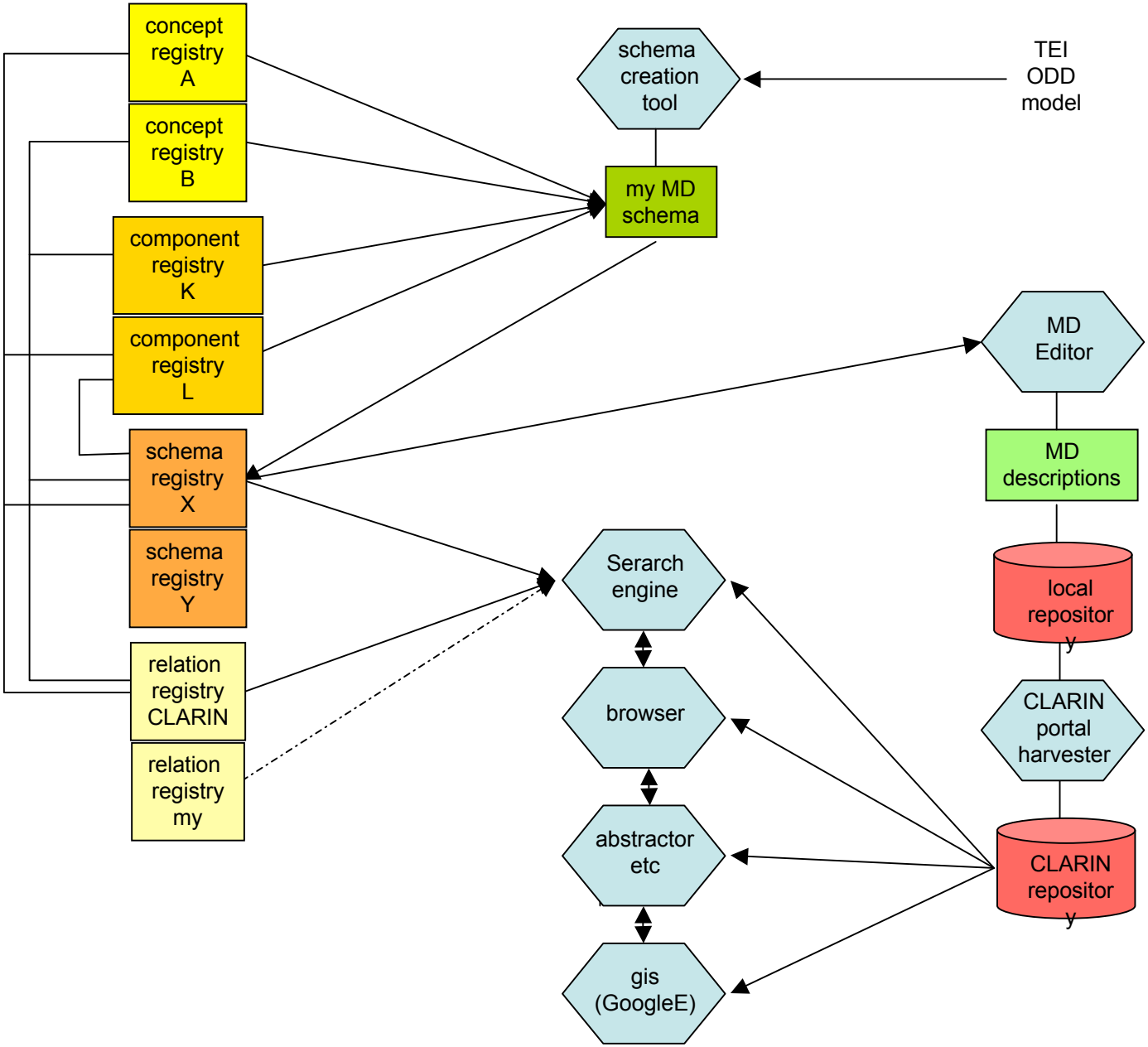
Perform search/browsing on the metadata catalog using the ISO DCR and other concept registries and CLARIN relation registry

Create metadata schema from selection of existing components. Allows creation of new components if they have references to ISOcat



The component registry in perspective.

The End


concept registry A
concept registry B
component registry K
component registry L
schema registry X
schema registry Y
relation registry CLARIN
relation registry my
schema creation tool
TEI ODD model
my MD schema
MD Editor
MD descriptions
local repository
CLARIN portal harvester
CLARIN repository
Serarch engine
browser
abstractor etc
gis (GoogleE)